Query and Visualization

Progress & Plans

Luis Sarmento

Goals of this Task

- development of tools for
  - querying extracted information
  - visualizing annotated documents and datasets
  - customize open source, tools and information to be used in the project as a set of web services / user interfaces
- Also, SAPO is supposed to provide / help with:
  - continuous scanning of the social web, news sources and various kinds of data streams
  - monitoring and information extraction of data streams

Data Streams

- News sources
  - National Level / Regional Level
  - About XX news items per day (national level)
- Twitter:
  - Crawling several thousands of PT users (some BR)
    - But we have been recently blacklisted from Twitter
    - We already have a reasonably large corpus of messages being used in several experiments

Data Streams

- User comments:
  - SAPO has its own platform for user comments being used in several popular news sites
    - <News Item, Comment, Community Votes>
    - Stats: Gamboa
- Blogs, Photos and Videos:
  - SAPO has also its own Bloggging, Photo and Video Sharing platform
- All this is already available for consulting, but…

Monitoring Streams

- We are starting to build a platform for *monitoring* large flows of data
- Goal:
  - have and integrated platform for accumulating and performing real-time querying of continuous flows of annotated or raw data:
  - Yesterday's news is no news
- So far:
  - Cluster Hadoop 20 machines being fed with log data
  - Pig Scripting for data crunching
  - Experimenting NoSQL database systems: Hbase

Interaction with ongoing projects

- “Twitter Channels”:
  - Filtering of twitter messages based on contents
    - Topic filtering (but “topic” may be an entity)
    - We are finishing a platform for creating channels on the fly:
      - Creating a training set, creating support lexicon, evaluating
- Voxx: project for mining quotations from news
  - Is evolving to a generic news mining project with several smaller task:
    - Acronym Mining
    - Short Biographical Description Mining

Short Biographical Descriptions

- We constantly mine news for certain structures:
  - The french prime-minister, Nicolas Sarkozy, …
  - We compile lists of tuples:
    - <name, bio_info, date, @news_titles, @tags_titles>
    - date, @news_titles, @tags_titles is used for time and topic disambiguation in later stages
  - We are finishing a service for answering “who is?” requests?
    - who _is (“Mourinho”, now(), tags = “sports”, keywords = “Real”)
    - Who_is (“french prime-mininster”, now())
- We can provide raw information to the POWER ontology

Next Steps

- Our major concern now is solving the technical problem
  - How to have continuous data flows coming into the system

    and simultaneously

    have “real-time” query capabilities (for visualization or exploration) over all data?
- We are also exploring several visualization packages
- We are trying to solve Twitter blacklisting problem.
- We are trying to exposes the data sources we have
- We are trying to make the CPU power available to partners (“Amazon style”)